

Article

Not peer-reviewed version

---

# MK-DCCA Based Fault Diagnosis for Incipient Fault in Nonlinear Dynamic Processes

---

[Junzhou Wu](#) , [Mei Zhang](#) <sup>\*</sup> , Lingxiao Chen

Posted Date: 14 September 2023

doi: 10.20944/preprints202309.0950.v1

Keywords: dynamic system; incipient fault; process monitoring; fault detection; MKPCA; DCCA



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Article

# MK-DCCA Based Fault Diagnosis for Incipient Fault in Nonlinear Dynamic Processes

Junzhou Wu , Mei Zhang \* and Lingxiao Chen

The Electrical Engineering College, Guizhou University, Guiyang 550000, China; wujunzhou959@gmail.com (J.W.); a1264481773@gmail.com (L.C.)

\* Correspondence: mzhang3@gzu.edu.cn; Tel.: +86-13385104650

**Abstract:** incipient fault diagnosis is particularly important in process industrial systems, as its early detection helps to prevent major accidents. Against this background, this study proposes a combined method of Mixed Kernel Principal Components Analysis and Dynamic Canonical Correlation Analysis (MK-DCCA). The robust generalization performance of this approach is demonstrated through experimental validation on a randomly generated dataset. Furthermore, Comparative experiments were conducted on a CSTR Simulink model, comparing the MK-DCCA method with DCCA and DCVA methods, demonstrating its excellent detection performance for incipient fault in nonlinear and dynamic system. Meanwhile, fault identification experiments were conducted, validating the high accuracy of fault identification method based on contribution. The experimental findings demonstrate that the method possesses a certain industrial significance and academic relevance.

**Keywords:** dynamic system; incipient fault; process monitoring; fault detection; MKPCA; DCCA

## 1. Introduction

As modern process industry systems evolve to become more complex, scaled, integrated, and intelligent, they often consist of numerous devices operating collaboratively, forming complex dynamic systems with multiple variables and significant time delays. The dynamic characteristics of these systems are progressively intricate, making the occurrence of faults inevitable. When a fault in any component of a system device goes unnoticed, the consequences encompass not only equipment damage but also the potential for degraded system performance, abnormal shutdowns, and even catastrophic consequences. Consequently, to uphold the reliability and safety of systems, and to guarantee the high-quality and efficient functioning of process industry systems, there is an urgent requirement to monitor, evaluate, and diagnose the real-time performance and operational state of all devices within the system. This is essential for implementing effective measures to ensure the stable operation of both the system and its components.

Existing research on fault diagnosis in process industries has been predominantly focused on the detection of abrupt faults. However, in recent times, both the industrial sector and the academic community have shown growing interest in detecting incipient faults. In fact, early detection of incipient faults is deemed even more significant than detecting abrupt faults. Hence, within process industry systems, the detection and localization of minor faults and the early stages of incipient fault development carry essential academic value and engineering significance. These efforts play a vital role in enabling effective fault remediation and ensuring the secure operation of the system.

To address the issue of diagnosing incipient faults in process industry systems, the existing approaches mainly fall into two categories: model-based methods and data-driven methods. Given that precise physical models are often unattainable for large-scale industrial processes [1], model-based methods encounter considerable constraints in their real-world implementation. Data-driven methods do not demand precise mechanistic models and are less dependent on process experiential knowledge, making them more suitable for extensive industrial processes. Common data-driven approaches are grounded in multivariate statistical analysis techniques, such as Principal Component Analysis (PCA),

Partial Least Squares (PLS), Canonical Variable Analysis (CVA), Canonical Variable Discriminative Analysis (CVDA), and Canonical Correlation Analysis (CCA). These methods have all proven their efficacy in industrial contexts [2]. The application of PCA-based methods has yielded favorable outcomes in the semiconductor manufacturing and aluminum smelting sectors [3–5]. Ding et al. utilized an enhanced PLS approach for forecasting and diagnosing key performance indicators in industrial hot-rolled strip steel mills [6], and similarly, a series of investigations have been conducted on PLS-based methods in [7]. In-depth research on the CVA method was undertaken by Ruiz-Cárcel et al. [8], while Pilario et al. proposed the CVDA method and its expanded iterations based on CVA [9–11]. [12] first employed data-driven CCA techniques to achieve residual generation based on canonical correlation, yielding favorable fault detection results. Subsequently, CCA-based methods have been extensively researched and improved by numerous scholars [13–17].

Nevertheless, the presence of dynamic behavior, nonlinearity, and other complex characteristics in industrial processes, coupled with the existence of closed-loop control strategies, renders the analysis and fault diagnosis of industrial processes even more challenging. Particularly when confronted with incipient faults, these characteristics significantly constrain the applicability of traditional multivariate statistical analysis methods. Despite the extensive efforts by researchers to investigate the diverse characteristics in industrial processes, the majority of research methods tend to concentrate on isolated characteristics rather than composite traits, avoiding the difficulties in the field of incipient fault diagnosis. For instance, [8,12] extended the CVA and CCA methods to dynamic versions, addressing the issue of process dynamics. However, they did not explore other characteristics, especially in the case of incipient faults, which could potentially impact their accuracy and applicability. [9] introduced an extended version of the CVA method called CVDA, along with the incorporation of Kernel Density Estimation (KDE) for calculating statistical indicator thresholds, which effectively addressing dynamic and non-Gaussian issues. Nevertheless, the problem of nonlinearity remained unresolved, and in practical industrial processes, incipient faults are often closely linked to the nonlinear behavior of systems. Although [10] proposed a combination of kernel methods and CVDA to tackle all characteristic issues, research on the relationship between nonlinear dynamic system inputs and outputs remains relatively limited. In the context of incipient faults, the consideration of the nonlinear relationships becomes especially crucial, as incipient faults can manifest as gradual changes in system behavior, where nonlinear characteristics may play a key role.

Building upon the aforementioned research foundation, in the face of the complex characteristics of high-dimensionality, nonlinearity, and dynamics associated with incipient faults in industrial processes, there is an urgent need for a novel multivariate statistical analysis approach to enhance the accuracy and reliability of diagnostics. This paper introduces a fault diagnosis method, called MK-DCCA and applies it to incipient fault diagnosis, aiming to achieve effective identification and accurate determination of incipient faults in industrial processes by considering multiple complex characteristics. Through this study, our intention is to offer novel perspectives and approaches to contribute to the ongoing development and real-world application of incipient fault diagnosis. The proposed method utilizes mixed kernel principal components analysis (MK-PCA) to map data into high-dimensional or even infinite-dimensional space to address nonlinear issues. The processed data is then employed as input for dynamic canonical correlation analysis (DCCA) to handle system dynamics in process monitoring. Lastly, a contribution-based approach is used for fault identification.

The subsequent sections of the paper are structured as follows: Section 2 provides an introduction to the fundamental theories of KPCA, DCCA, and the contribution-based fault recognition method; In Section 3, the MK-DCCA method used in this paper is proposed and thoroughly explicated; Section 4 employs two case studies to validate the effectiveness of the MK-DCCA method and the contribution-based fault recognition approach. In Case I, the proposed method is first applied to a randomly generated dataset to demonstrate its robust generalization performance. Subsequently, Case II utilizes the method on a simulated model of a continuous stirred tank reactor (CSTR). Comparative

experiments are conducted against various versions of CVA and CCA methods, demonstrating the favorable performance of the method in process monitoring and fault diagnosis.

## 2. Methodological Theory

### 2.1. Kernel Principal Component Analysis

KPCA employs kernel techniques to map data into a high-dimensional feature space, enabling original data to be linearly separable or approximately linearly separable in the new space. In detail, for nonlinear data matrix  $X$ , a nonlinear mapping is first employed to map all samples in  $X$  to a high-dimensional or even infinite-dimensional space (i.e., feature space), making them linearly separable. Subsequently, PCA dimensionality reduction is performed in this high-dimensional space.

Based on the method proposed by Schoölkopf et al. [18], the initial step consists of applying a kernel function to calculate the kernel matrix  $K$  using the following formula:

$$K(x_i, x_j) \triangleq K_{ij} = \langle \Phi(x_i), \Phi(x_j) \rangle \quad (1)$$

Where  $x_i$  and  $x_j$  ( $i, j = 1, 2, \dots, N$ ) represent the  $i$ th and  $j$ th samples,  $\Phi(\cdot)$  denotes the kernel function. The commonly employed kernel functions consist of polynomial kernel functions and Gaussian kernel functions (RBF), with their expressions presented as follows:

$$\begin{aligned} K_{poly}(x, x') &= (\langle x, x' \rangle + 1)^d \\ K_{RBF}(x, x') &= \exp\left(-\frac{\|x - x'\|^2}{c}\right) \end{aligned} \quad (2)$$

Where  $K_{poly}$  represents the polynomial kernel function,  $d$  is the parameter indicating the polynomial degree. This kernel satisfies the Mercer condition for  $d \in \mathbb{N}$  [19],  $c$  denotes the kernel width, which satisfies the Mercer condition for  $c > 0$  [20].

After obtaining a  $N \times N$  symmetric kernel matrix  $K$ , it is necessary to perform centering on it. The specific calculation formula is as follows:

$$\hat{K} = K - 1_N K - K 1_N + 1_N K 1_N \quad (3)$$

Where  $1_N \in \mathbb{R}^{N \times N}$  and  $(1_N)_{ij} = 1/N$ . Subsequently, PCA dimensionality reduction is performed on the centered kernel matrix  $\hat{K}$ . According to the equation below, perform an eigenvalue decomposition on  $\hat{K}$ .

$$\hat{K}v = N\lambda v \quad (4)$$

Where  $v$  represents an eigenvector, and  $\lambda$  denotes an eigenvalue. Subsequently,  $\hat{K}$  is diagonalized as

$$\hat{K}/N = S\Lambda S^T \quad (5)$$

Here,  $S = [v_1, v_2, \dots, v_N] \in \mathbb{R}^{N \times N}$  represents  $N$  eigenvectors, and  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_N) \in \mathbb{R}^{N \times N}$  denotes eigenvalues, where  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N$ . To preserve relevant information, the first  $r$  principal components are selected to explain 85% of the total variance. Subsequently, the kernel principal components  $t_k$  are calculated through the following projection:

$$T = [t_k] = S_r^T \hat{K} \in \mathbb{R}^{r \times N} \quad (6)$$

Where  $S_r$  denotes the first  $r$  columns of the eigenvector matrix  $S$ .

After processing the training set, any test data  $x_k^{test}$  at the  $k$ th sampling time is standardized using the mean and standard deviation of the training set to obtain  $\hat{x}_k^{test}$ . Afterward, the constructed kernel mapping from earlier is used to project it into the feature space according to the following formula:

$$k_k^{test} = (\hat{x}_k^{test}, \hat{x}_j) \in \mathbb{R}^{1 \times N} \quad (7)$$

Where  $\hat{x}_j$  represents all training samples,  $j = 1, \dots, N$ , Then  $k_k^{test}$  is centered by subtracting as

$$\hat{k}_k^{test} = k_k^{test} - 1_N^{test} K - k_k^{test} 1_N + 1_N^{test} K 1_N \quad (8)$$

Here  $1_N^{test} \in \mathbb{R}^{1 \times N}$  and  $(1_N^{test})_{ij} = 1/N$ . Ultimately, the kernel principal components of the test data at the  $k$ th sampling time are computed using the following formula:

$$t_k^{test} = S_r^T (\hat{k}_k^{test})^T \in \mathbb{R}^r \quad (9)$$

## 2.2. Dynamic Canonical Correlation Analysis

Given the limitations of methods such as CVA and CVDA, which are unable to adequately and effectively exploring the relationship between system input and output variables, this research opts for CCA as the cornerstone of the process monitoring approach, further extending its capabilities. Given the premise that the considered dynamic process is linear time-invariant, we assume that the process has process white noise and measurement white noise, and can be represented by a standard model described by a state space. Its mathematical expression is as follows:

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) + w(k) \\ y(k) &= Cx(k) + Du(k) + v(k) \end{aligned} \quad (10)$$

Where  $x \in \mathbb{R}^n$  is the state vector,  $u \in \mathbb{R}^l$  and  $y \in \mathbb{R}^m$  are input and output vectors, and  $w \in \mathbb{R}^n$  and  $v \in \mathbb{R}^m$  denote process and measurement noises, respectively. Matrix  $A$ ,  $B$ ,  $C$ , and  $D$  are unknown constant matrices with appropriate dimensions. In this study, it is further assumed that the process is stable. Under steady-state conditions, it holds that:  $\lim_{k \rightarrow \infty} \mu_x(k) = \mu_x$  and  $\lim_{k \rightarrow \infty} \Sigma_x(k) = \Sigma_x$ , where  $\mu_x$  and  $\Sigma_x$  are constants. Therefore, the cross-covariance between input and output remains constant.

In [12], the concept of DCCA was first introduced as an extension of the CCA-based methodology, employed for detecting faults in such dynamic systems under steady-state conditions. Leveraging the stochastic system model (10), an investigation is conducted into the dependency of the future output  $y_f$  on past input, past output  $z_p$ , and future input  $u_f$ . To achieve this, firstly, data structures and sets are defined, assuming  $p$  and  $f$  to be lag and lead parameters. The lagged variables and their corresponding data matrices are defined as follows.

$$z_p(k) = \begin{bmatrix} y(k-p) \\ \vdots \\ y(k-1) \\ u(k-p) \\ \vdots \\ u(k-1) \end{bmatrix}; y_f(k) = \begin{bmatrix} y(k) \\ \vdots \\ y(k+f) \end{bmatrix}; u_f(k) = \begin{bmatrix} u(k) \\ \vdots \\ u(k+f) \end{bmatrix} \quad (11)$$

$$\begin{aligned} Z_p &= [z_p(1), \dots, z_p(N)] \in \mathbb{R}^{(s(m+l) \times N)} \\ Y_f &= [y_f(1), \dots, y_f(N)] \in \mathbb{R}^{((f+1)m \times N)} \\ U_f &= [u_f(1), \dots, u_f(N)] \in \mathbb{R}^{((f+1)l \times N)} \end{aligned} \quad (12)$$

Qin et al. demonstrated that equation (10) can be reformulated as:

$$\begin{aligned} x(k+1) &= A_K x(k) + B_K u(k) + Ky(k) \\ y(k) &= Cx(k) + Du(k) + e(k) \end{aligned} \quad (13)$$

Where  $A_K = A - KC$ ,  $B_K = B - KD$ , with  $K$  serving as the Kalman filter gain matrix, ensuring that the eigenvalues of  $A_K$  are situated within the unit circle to guarantee system stability, and  $e(k)$  represents the innovation sequence. It is evident from equation (13) that the following equations hold:

$$x(k) = A_K^s x(k-s) + \sum_{i=1}^s A_K^{i-1} \begin{bmatrix} K & B_K \end{bmatrix} \begin{bmatrix} y(k-i) \\ u(k-i) \end{bmatrix} \quad (14)$$

$A_K$  is stable, and simultaneously, selecting a sufficiently large value for  $s$  results in  $A_K^s \approx 0$ , subsequently,

$$x(k) \approx P^T z_p(k) \quad (15)$$

Where  $P^T = \begin{bmatrix} P_y & P_u \end{bmatrix}$ ,  $P_y = \begin{bmatrix} A_K^{s-1}K & \dots & A_K K & K \end{bmatrix}$ ,  $P_u = \begin{bmatrix} A_K^{s-1}B_K & \dots & A_K B_K & K \end{bmatrix}$ . The past measured value  $z_p(k)$  encompasses process input and output data within the time interval  $[k-s, k-1]$ , as shown in equation (12). Additionally, according to equations (13), the follow equation is also valid:

$$y_f(k) = \Gamma_{K,f}(k) + H_{K,u,f}u_f(k) + H_{K,y,f}y_f(k) + e_f(k) \quad (16)$$

Here,

$$\begin{aligned} \Gamma_{K,f} &= \begin{bmatrix} C \\ CA_K \\ \vdots \\ CA_K^{f-1} \end{bmatrix} & H_{K,u,f} &= \begin{bmatrix} D & 0 & \dots & 0 \\ CB_K & D & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ CA_K^{f-1}B_K & \dots & CB_K & D \end{bmatrix} \\ e_f(k) &= \begin{bmatrix} e(k) \\ e(k+1) \\ \vdots \\ e(k+f) \end{bmatrix} & H_{K,y,f} &= \begin{bmatrix} 0 & 0 & \dots & 0 \\ CK & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ CA_K^{f-1}K & \dots & CK & 0 \end{bmatrix} \end{aligned} \quad (17)$$

According to equation(15),

$$\begin{aligned} (I - H_{K,y,f})y_f(k) &\approx \Gamma_{K,f}P^T z_p(k) + H_{K,u,f}u_f(k) + e_f(k) \\ &= \begin{bmatrix} \Gamma_{K,f}P^T H_{K,u,f} \end{bmatrix} \begin{bmatrix} z_p(k) \\ u_f(k) \end{bmatrix} + e_f(k) \end{aligned} \quad (18)$$

Formula (18) can be further written as:

$$L^T y_f(k) = M^T \begin{bmatrix} z_p(k) \\ u_f(k) \end{bmatrix} + e_f(k) \quad (19)$$

Here,  $L = (I - H_{K,y,s_f})^T$ ,  $M = \begin{bmatrix} \Gamma_{K,s_f}P^T H_{K,u,s_f} \end{bmatrix}^T$



Subsequently, by employing CCA technique in residual generation, the issue of fault detection in dynamic processes is resolved. Process input and output data are structured based on time intervals, denoted as  $Y_f$  and  $\begin{bmatrix} Z_p \\ U_f \end{bmatrix}$ . Centralize  $Y_f$  and  $\begin{bmatrix} Z_p \\ U_f \end{bmatrix}$ , and then

$$\begin{bmatrix} \Sigma_z & \Sigma_{z,y_f} \\ \Sigma_{y_f,z} & \Sigma_{y_f} \end{bmatrix} \approx \frac{1}{N-1} \begin{pmatrix} \begin{bmatrix} Z_p \\ U_f \end{bmatrix} \begin{bmatrix} Z_p \\ U_f \end{bmatrix}^T & \begin{bmatrix} Z_p \\ U_f \end{bmatrix} Y_f^T \\ Y_f \begin{bmatrix} Z_p \\ U_f \end{bmatrix}^T & Y_f Y_f^T \end{pmatrix} \quad (20)$$

Through the utilization of CCA, the weighting matrices  $J_d$  and  $L_d$  can be obtained from the subsequent equations.

$$\Sigma_z^{-1/2} \Sigma_{z,y_f} \Sigma_{y_f}^{-1/2} = \Gamma \Lambda \Delta^T; J_d = \Sigma_z^{-1/2} \Gamma(:, 1:n); L_d = \Sigma_{y_f}^{-1/2} \Delta(:, 1:n) \quad (21)$$

$$\Lambda = \begin{bmatrix} \Lambda_l & 0 \\ 0 & 0 \end{bmatrix}$$

Where  $\Lambda_n = \text{diag}(\lambda_1, \dots, \lambda_n)$ . The Cumulative Percentage Value (CPV) method can be utilized to determine the system order  $n$  [21]. It is important to highlight that

$$J_d^T \begin{bmatrix} Z_p \\ U_f \end{bmatrix} \begin{bmatrix} Z_p \\ U_f \end{bmatrix}^T J_d = I; L_d^T Y_f Y_f^T L_d = I \quad (22)$$

The following equations can be derived from formula (21):

$$J_d^T \begin{bmatrix} Z_p \\ U_f \end{bmatrix} Y_f^T L_d = \Lambda_n \quad (23)$$

It is reasonable to define the residual vector as presented below based on formula (23),

$$r(k) = L_d^T y_f(k) - M_d^T \begin{bmatrix} z_p(k) \\ u_f(k) \end{bmatrix} \quad (24)$$

Where  $M_d^T = \Lambda_n J_d^T$ . Furthermore, the covariance matrix of  $r(k)$  can be estimated as:

$$\begin{aligned} & \left( L_d^T Y_f - M_d^T \begin{bmatrix} Z_p \\ U_f \end{bmatrix} \right) \left( L_d^T Y_f - M_d^T \begin{bmatrix} Z_p \\ U_f \end{bmatrix} \right)^T \\ &= L_d^T Y_f Y_f^T L_d + \Lambda_n^2 J_d^T \begin{bmatrix} Z_p \\ U_f \end{bmatrix} \begin{bmatrix} Z_p \\ U_f \end{bmatrix}^T - 2\Lambda_n J_d^T \begin{bmatrix} Z_p \\ U_f \end{bmatrix} Y_f^T L_d \\ &= I - \Lambda_n^2 \end{aligned} \quad (25)$$

Residuals of canonical variables follow a multivariate normal distribution with zero mean, and the covariance matrix is given by equation (25). Therefore, it is reasonable to utilize the following statistical data for detection purposes:

$$T_r^2(k) = (N-1) r^T(k) (I - \Lambda_n^2)^{-1} r(k) \quad (26)$$

The threshold  $J_{th,T^2}$  can be defined as:

$$J_{th,T^2} = \frac{n(N^2 - n)}{N(N - n)} \mathcal{F}_{1-\alpha}(n, N - n) \quad (27)$$

### 2.3. Contribution based Fault Identification

Building upon the research by Li et al., this research conducted a comparison of fault identification capabilities among the traditional  $Q$  contribution method, the  $T^2$  contribution method, and the contribution method based on residuals of canonical variables [22]. In this section, the samples are consolidated into a dataset  $Y$ , rather than being partitioned into input and output categories. Initially, lag parameter  $p$  and lead parameter  $f$  are introduced, followed by a redefinition of data structure and sets to derive past observation vectors and future observation vectors.

$$y_{p,k} = \begin{bmatrix} y_{k-1} \\ y_{k-2} \\ \vdots \\ y_{k-p} \end{bmatrix} \in \mathbb{R}^{np}; y_{f,k} = \begin{bmatrix} y_k \\ y_{k+1} \\ \vdots \\ y_{k+f-1} \end{bmatrix} \in \mathbb{R}^{nf} \quad (28)$$

Here,  $y_k$  represents the  $k$ -th sample, and  $n$  denotes the number of variables included in each sample. To prevent variables with large values from dominating, normalization of  $y_{p,k}$  and  $y_{f,k}$  is required. Subsequently, the rearranged normalized past and future observation vectors,  $\hat{y}_{p,k}$  and  $\hat{y}_{f,k}$ , are presented as follows:

$$\begin{aligned} \hat{Y}_p &= [\hat{y}_{p,k+1}, \hat{y}_{p,k+2}, \dots, \hat{y}_{p,k+M}] \in \mathbb{R}^{np \times M} \\ \hat{Y}_f &= [\hat{y}_{f,k+1}, \hat{y}_{f,k+2}, \dots, \hat{y}_{f,k+M}] \in \mathbb{R}^{nf \times M} \end{aligned} \quad (29)$$

Here,  $M = N - p - f + 1$ , where  $N$  denotes the number of samples. Then the covariance matrices of  $\hat{y}_p$  and  $\hat{y}_f$  can be computed using the following formulas:

$$\begin{aligned} \Sigma_{pp} &= \frac{1}{N-1} \hat{Y}_p \hat{Y}_p^T \\ \Sigma_{ff} &= \frac{1}{N-1} \hat{Y}_f \hat{Y}_f^T \\ \Sigma_{fp} &= \frac{1}{N-1} \hat{Y}_f \hat{Y}_p^T \end{aligned} \quad (30)$$

Subsequently, performing singular value decomposition on the Hankel matrix  $H$  yields the following results:

$$H = \Sigma_{ff}^{-1/2} \Sigma_{fp} \Sigma_{pp}^{-1/2} = U \Sigma V^T \quad (31)$$

Where  $U = (u_1, \dots, u_l)$ ,  $V = (v_1, \dots, v_m)$ ,  $\Sigma = \begin{bmatrix} \Sigma_q & 0 \\ 0 & 0 \end{bmatrix}$ ,  $u_i$  and  $v_j$  are the corresponding singular vectors,  $\Sigma_q = \text{diag}(\lambda_1, \dots, \lambda_q)$ ,  $\lambda_1 \geq \lambda_2, \dots, \geq \lambda_q \geq 0$ , are the singular values. The value of  $k$  can be determined using the CPV method.

#### 2.3.1. Q-based Contribution

The canonical residual variable  $e_k$ , employed for contribution calculation, can be obtained from the subsequent formula:

$$e_k = G \hat{y}_{p,k} = V_{np-p}^T \Sigma_{pp}^{-1/2} \hat{y}_{p,k} \quad (32)$$



Following the definition of variable contributions based on CVA proposed by Jiang et al.[23], the calculation of variable contributions using the  $Q$  statistical metric is presented as follows:

$$C_Q = Q = e^T e = e^T G \hat{y}_{p,k} = \sum_{i=1}^n \sum_{j=1}^{np-q} e_j G_{j,i} \hat{y}_{p,i} = \sum_{i=1}^n C_{i,Q} \quad (33)$$

Where  $C_{i,Q}$  is the contribution of variable  $\hat{y}_i$  to the monitoring statistic  $Q$ , and  $e_j G_{j,i} \hat{y}_{p,i}$  signifies the contribution of variable  $\hat{y}_i$  to the  $j$ -th canonical residual variable  $e_j$ . Ultimately, by dividing each variable's contribution of  $\hat{y}_i$  to  $Q$  by the cumulative contribution  $C_Q$ , the percentage of each contribution can be determined, thus identifying the variables associated with faults.

$$P_{i,Q} = \frac{C_{i,Q}}{C_Q} \quad (34)$$

### 2.3.2. $T^2$ -based Contribution

Also following the CVA method, the calculation formula for the canonical state variable  $z_k$ , used in contribution assessment, is provided below:

$$z_k = K \hat{y}_{p,k} = V_q^T \Sigma_{pp}^{-1/2} \hat{y}_{p,k} \quad (35)$$

Additionally, in accordance with [22], the computation of variable contribution based on the  $T^2$  statistical indicator can be expressed as:

$$C_{T^2} = T^2 = z^T z = z^T K \hat{y}_{p,k} = \sum_{i=1}^n \sum_{j=1}^q z_j K_{j,i} \hat{y}_{p,i} = \sum_{i=1}^n C_{i,T^2} \quad (36)$$

Where  $C_{i,T^2}$  signifies the contribution of variable  $\hat{y}_i$  to the monitoring statistic  $T^2$ , and  $z_j K_{j,i} \hat{y}_{p,i}$  denotes the contribution of variable  $\hat{y}_i$  to the  $j$ -th typical state variable  $z_j$ . Ultimately, the percentages of each contribution can be computed by dividing the contribution of each variable  $\hat{y}_i$  to  $T^2$  by the cumulative contribution  $C_{T^2}$ , facilitating the identification of variables correlated with faults.

$$P_{i,T^2} = \frac{C_{i,T^2}}{C_{T^2}} \quad (37)$$

### 2.3.3. $T_d$ -based Contribution

Apart from the two contribution calculation methods mentioned above, [22] also introduced a contribution calculation approach based on Canonical Variable Residuals (CVR). The central idea is to detect minor changes by examining the deviations between future and past canonical variables. The definition of canonical variable residuals is provided below:

$$r_k = L_q^T \hat{y}_{f,k} - \Sigma_q J_q^T \hat{y}_{p,k} \quad (38)$$

Where  $L_q^T$  denotes the first  $q$  rows of the matrix  $L^T$ , and  $L^T = U_q^T \Sigma_{ff}^{-1/2}$ . Similarly,  $J_q^T$  represents the first  $q$  rows of the matrix  $J^T$ , and  $J^T = V_q^T \Sigma_{pp}^{-1/2}$ .  $\Sigma_q$  represents a diagonal matrix composed of the first  $q$  singular values. The calculation of variable contributions using the  $T_d$  statistical metric based on CVR is presented below:

$$\begin{aligned} C_{T_d} &= T_d = r_k^T (I - \Sigma_q^2)^{-1} r_k = r_k^T (I - \Sigma_q^2)^{-1} (L_q \hat{y}_{f,k} - \Sigma_q J_q \hat{y}_{p,k}) \\ &= \sum_{i=1}^n \sum_{j=1}^q r_j \Sigma_{dd_j}^{-1} (L_{j,i} \hat{y}_{f,i} - \Sigma_j J_{j,i} \hat{y}_{p,i}) = \sum_{j=1}^n C_{i,T_d} \end{aligned} \quad (39)$$

Where  $C_{i,T_d}$  denotes the contribution of variable  $\hat{y}_i$  to the monitoring statistic  $T_d$ ,  $\Sigma_j$  represents the  $j$ -th singular value, and  $\Sigma_{dd_j}^{-1}$  is the  $j$ -th diagonal element of the matrix  $(I - \Sigma_q^2)^{-1}$ . Ultimately, the percentages of each contribution can be computed by dividing the contribution of each variable  $\hat{y}_i$  to  $T_d$  by the cumulative contribution  $C_{T_d}$ , aiding in identifying variables correlated with faults.

$$P_{i,T_d} = \frac{C_{i,T_d}}{C_{T_d}} \quad (40)$$

### 3. MK-DCCA Method

In this section, building upon the aforementioned theoretical foundation, the MK-DCCA method utilized in this study is introduced. To endow the kernel function with both strong interpolation and extrapolation capability, ensuring a robust generalization performance, a combination scheme of local and global kernels is chosen based on the research foundation of KPCA. This involves combining the RBF kernel and polynomial kernel to form a mixture kernel. The detailed formulas are as follows:

$$K_{mix} = \omega K_{poly} + (1 - \omega) K_{RBF} \quad (41)$$

Here,  $\omega \in [0, 1]$  represents the mixing weight, and the mixture kernel reverts to the polynomial kernel and RBF kernel when  $\omega = 1$  and  $\omega = 0$  respectively. [24] proposes utilizing a weighted sum of linear ( $d = 1$ ) and RBF kernels to balance favorable interpolation and extrapolation capabilities. Accordingly, this study employs a combination of a polynomial kernel with  $d = 1$  and an RBF kernel.

Subsequently, the data processed by the MKPCA method is employed for performing the DCCA method. Firstly, the input vector  $u$  and output vector  $y$  are standardized to obtain  $\hat{u}$  and  $\hat{y}$ . Subsequently, the data structure and sets are established, assuming  $p$  and  $f$  as lag and lead parameters, and the past and future observation vectors of  $u$  and  $y$  are defined as

$$\begin{aligned} u_p(k) &= \begin{bmatrix} u(k-p) \\ \vdots \\ u(k-1) \end{bmatrix}; \quad y_p(k) = \begin{bmatrix} y(k-p) \\ \vdots \\ y(k-1) \end{bmatrix} \\ u_f(k) &= \begin{bmatrix} u(k) \\ \vdots \\ u(k+f) \end{bmatrix}; \quad y_f(k) = \begin{bmatrix} y(k) \\ \vdots \\ y(k+f) \end{bmatrix} \end{aligned} \quad (42)$$

Furthermore, the new past and future observation matrices are defined as

$$\begin{aligned} Z_p &= [z_p(1), \dots, z_p(N)] \in \mathbb{R}^{p(m+l) \times N} \\ Y_f &= [y_f(1), \dots, y_f(N)] \in \mathbb{R}^{(f+1)l \times N} \\ z_p(k) &= \begin{bmatrix} y_p(k) \\ u_p(k) \end{bmatrix}; \quad Z = \begin{bmatrix} Z_p \\ U_f \end{bmatrix} \end{aligned} \quad (43)$$

Next,  $Z$  will be used as the input matrix and  $Y_f$  as the output matrix to perform a CCA-based process monitoring method. Initially, compute the self-covariance and cross-covariance matrices  $\Sigma_{zz}$ ,  $\Sigma_{y_f y_f}$ , and  $\Sigma_{zy_f}$  for  $Z$  and  $Y_f$  using equation (45).

$$\begin{aligned}\Sigma_{zz} &= \frac{1}{N-1} \sum_{i=1}^N \hat{z}(i) \hat{z}^T(i) \\ \Sigma_{y_f y_f} &= \frac{1}{N-1} \sum_{i=1}^N \hat{y}_f(i) \hat{y}_f^T(i) \\ \Sigma_{zy_f} &= \frac{1}{N-1} \sum_{i=1}^N \hat{z}(i) \hat{y}_f^T(i)\end{aligned}\quad (44)$$

Where  $N$  is the number of samples,  $\hat{u}(i)$  and  $\hat{y}(i)$  denote the normalized input and output samples at the  $i$ -th time instance, respectively. Subsequently, the Hankel matrix  $H$  is constructed using the subsequent formula:

$$H = \Sigma_{zz}^{-1/2} \Sigma_{zy_f} \Sigma_{y_f y_f}^{-1/2} \quad (45)$$

Through singular value decomposition, the matrix  $H$  can be decomposed into

$$H = \Gamma \Lambda \Delta^T \quad (46)$$

Where  $\Gamma = (\gamma_1, \dots, \gamma_l)$ ,  $\Delta = (\delta_1, \dots, \delta_m)$ ,  $\Lambda = \begin{bmatrix} \Lambda_k & 0 \\ 0 & 0 \end{bmatrix}$   $\gamma_i$  and  $\delta_j$  are the corresponding singular vectors,  $\Lambda_k = \text{diag}(\lambda_1, \dots, \lambda_k)$ ,  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k \geq 0$  represents the singular values. Based on equation (24), obtaining the unknown constant matrices  $L$  and  $M$  is all that is required to derive the residual signal. Let

$$\begin{aligned}L_n &= \Sigma_{y_f y_f}^{-1/2} \Delta(:, 1:n) \\ J_n &= \Sigma_{zz}^{-1/2} \Gamma(:, 1:n) \\ M_n^T &= \Lambda_n J_n^T\end{aligned}\quad (47)$$

Moreover, the covariance matrix of the residual signal  $r(k)$  can be estimated as

$$\Sigma_{rr} = I - \Lambda_n^2 \quad (48)$$

Where  $I$  denotes the identity matrix. Finally, the statistical metric  $T^2$  utilized for dynamic system process monitoring can be calculated through the subsequent equation:

$$T^2(k) = (N-1) r^T(k) \Sigma_{rr}^{-1} r(k) \quad (49)$$

The calculation of the corresponding threshold is as follows:

$$T_{th}^2 = \frac{n(N^2 - n)}{N(N - n)} \mathcal{F}_{1-\alpha}(n, N - n) \quad (50)$$

Here,  $n$  represents the chosen number of singular values, and  $N$  is the number of samples. After obtaining the threshold, process monitoring is conducted according to the subsequent logic:

If  $T^2 > T_{th}^2$ , it indicates a fault; otherwise, there is no fault.

Ultimately, if a fault is detected, the contribution-based fault identification method from Section 2.3 is utilized to identify the fault variables and achieve accurate fault localization.

#### 4. Case Study

In this section, the experimental validation of the proposed method is conducted through two case studies. Firstly, experimental analysis is conducted on a randomly generated dataset to verify the

generalization performance of the proposed method. Subsequently, experimental analysis is conducted on a CSTR simulation model, with a comparison made against CVA and CCA methods, along with their improved versions, to demonstrate the superiority of the proposed approach.

4.1. Case I: Case study using randomly generated data

4.1.1. Model Introduction

In this subsection, the process model utilized for data generation is defined as follows:

$$\begin{aligned} u(k) &= Wx(k) + e(k) \\ y(k) &= \Phi u(k) + b + v(k) \end{aligned}$$

(51)

Where  $u(k)$  and  $y(k)$  represent the input and output, respectively, and  $W$ ,  $b$ , and  $\Phi$  are constants. Initially, the model is employed to generate a training dataset containing 2000 samples with 6 features each. Subsequently, test datasets are created for actuator faults, sensor faults, and process faults, and the details of dataset are provided in Table 1.

Table 1. The information of datasets.

Fault Index	Fault Location	Fault Category	Fault Variables	Sample Count	Feature Count	Introduction Time(s)
Fault-Free	/	/	/	2000	6	/
Fault 1	Sensor	Incipient	9th	2000	6	1000
Fault 2	Actuator	Abrupt	1st	2000	6	1000

4.1.2. Process Monitoring

In this subsection, the fault-free dataset generated in Section 4.1.1 is employed as the training set to train the model. The fault datasets are then used as the test set for process monitoring to demonstrate the performance of the proposed method. The monitoring results for Fault 1 and Fault 2 are shown in Figure 1, separately.

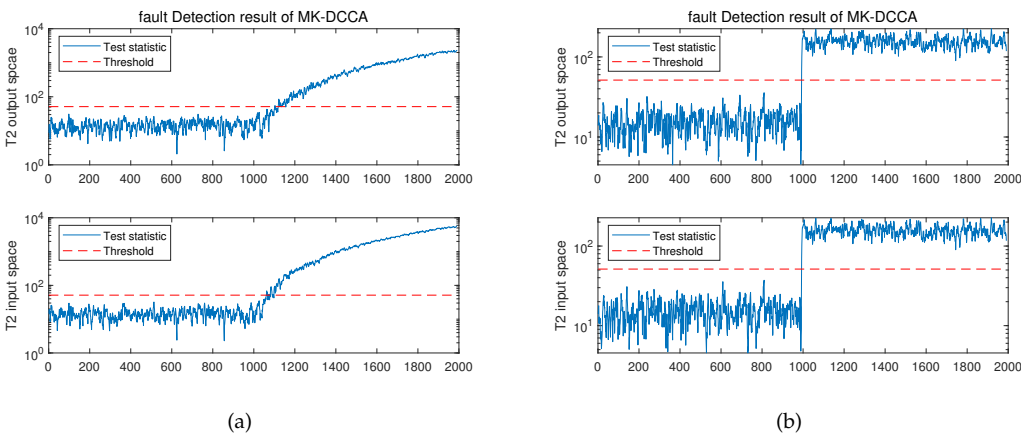


Figure 1. Monitoring Results of the Randomly Generated Dataset by MK-DCCA :(a)Monitoring Results for Fault 1.(b)Monitoring Results for Fault 2.

The monitoring graphs reveal that the proposed approach effectively identifies the abnormal states of the system and provides early warnings. For Fault 1 (incipient fault), the monitoring model can provide an alert at the 54th sample after fault occurs and effectively forecast evolving trend of faults. Moreover, for another prevalent fault, namely Fault 2 (abrupt fault), the monitoring model

can promptly alert about abnormal operating state of system as soon as fault occurs. Simultaneously, the proposed approach exhibits satisfactory performance during the monitoring processes of both aforementioned fault types. To comprehensively evaluate performance of the method, this study quantifies it using four metrics: Fault Detection Rate (FDR), False Alarm Rate (FAR), Miss Detection Rate (MDR), and Fault Detection Time (FDT). FDT indicates the time when monitoring model initiates the first alert after a fault occurs. The definitions for the other three metrics are provided below:

$$FDR = \frac{TP}{TP+FN}$$
$$FAR = \frac{FP}{FP+TN}$$
$$MDR = \frac{FN}{TP+FN}$$

(52)

Where  $TP$  is the number of faults correctly detected,  $TN$  represents the number of normals correctly detected,  $FP$  denotes the number of normal samples incorrectly reported as faults, and  $FN$  signifies the number of faults incorrectly reported as normals. Ultimately, utilizing the aforementioned metrics, the quantified results of monitoring performance are displayed in Table 2.

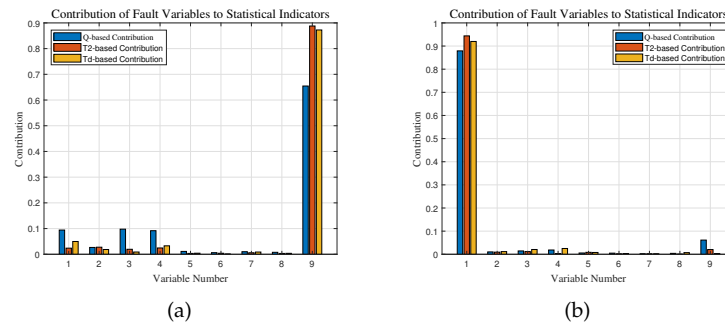
Table 2. Monitoring Performance Indicators of Fault 1 and Fault 2.

Fault Index	Statistical Metrics	FDR(%)	FAR(%)	MDR(%)	FDT(s)
Fault 1	$T_{in}^2$	91.9	0	8.1	1054
	$T_{out}^2$	92.5	0	7.5	1054
Fault 2	$T_{in}^2$	99.9	0	0.1	1000
	$T_{out}^2$	99.9	0.1	0.1	1000

It is evident that proposed approach achieves a fault detection rate of 92.5% for the incipient fault (Fault 1) and attains a higher 99.9% detection rate for the comparatively easily detectable abrupt fault (Fault 2). As for the false alarm rate, MK-DCCA method maintains an excellent performance of zero false alarms for both fault types. Meanwhile, proposed method successfully manages to maintain the miss detection rate at an acceptable level of 7.5% for Fault 1, whereas for the more readily detectable Fault 2, the miss detection rate decreases to 0.1%. Ultimately, the fault detection time for Fault 1 occurs at the 1054th sample (54 samples following the fault occurrence), while for Fault 2, it is the 1000th sample (immediately following the fault occurrence). The reason behind this phenomenon is that incipient faults exhibit a relatively minor impact on monitoring indicators during the early stages of fault occurrence, necessitating a certain degree of fault development to trigger model warnings.

4.1.3. Fault Identification

This section of the experiment primarily aims to evaluate the fault identification capability of proposed method. It involves identify the fault variables by evaluating the contribution of each variable to the fault detection indicators, with the goal of achieving accurate fault localization. The variable contribution plots of Fault 1 and Fault 2 are illustrated in Figure 2:



**Figure 2.** Contribution Plot of Process Variables under Fault Scenarios:(a)Contribution Plot of Fault 1.(b)Contribution Plot of Fault 2.

Table 1 indicates that the fault variable for Fault 1 is the 9th variable, while for Fault 2, it is the 1st variable. This conclusion is also apparent from Figure 2. It is evident that in both types of fault identification experiments, the  $T^2$  indicator contribution performed the best. For Fault 1, variable 9 contributed 96.7% to the fault statistic indicator, while for Fault 2, variable 1 contributed 98.6%. Hence, the fault identification method employed in this research demonstrates a satisfactory level of accuracy.

In conclusion, the process monitoring and fault identification experiments conducted on a randomly generated dataset from a standard process model have demonstrated strong generalization performance of proposed method. Simultaneously, it has exhibited satisfactory monitoring performance and fault identification accuracy.

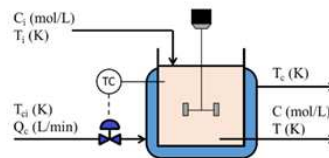
#### 4.2. Case II: Case Analysis of CSTR Simulation Model

##### 4.2.1. Model Introduction

The dataset employed in this case study is generated by a CSTR Simulink simulation model tailored for simulating incipient faults. A detailed description of the model can be found in [9]. The schematic diagram of the CSTR model is shown in Figure 3, and Table 3 summarizes all the process variables of the system. The system inputs are  $C_i$ ,  $T_i$ , and  $T_{ci}$ , while the system outputs are  $C$ ,  $T$ ,  $T_c$ , and  $Q_c$ . The dynamic model of CSTR process is described as follows:

$$\begin{aligned} \frac{dC}{dt} &= \frac{Q}{V} (C_i - c) - a_1 k C + v_1 \\ \frac{dT}{dt} &= \frac{Q}{V} (T_i - T) - a_1 \frac{(\Delta H_r) k C}{\rho C_p V} (T - T_c) + v_2 \\ \frac{dT_c}{dt} &= \frac{Q_c}{V_c} (T_{ci} - T_c) + b_1 \frac{UA}{\rho C_{pc} V_c} (T - T_c) + v_3 \end{aligned} \quad (53)$$

Where  $Q$  is the inlet flow rate,  $\Delta H_r$  is the heat of reaction,  $UA$  is the heat transfer coefficient,  $\rho$  and  $\rho_c$  are the fluid density,  $C_p$  and  $C_{pc}$  are the heat capacity of the fluid, and  $V$  and  $V_c$  are the volumes of the tank and jacket, respectively.



**Figure 3.** Schematic Diagram of the CSTR Model.

Table 3. Process Variables Involved in the CSTR System.

Variable Index	1	2	3	4	5	6	7	8	9	10
Variable Name	$C_i$	$T_i$	$T_{ci}$	$C_i$	$T_i$	$C$	$T$	$T_c$	$T_{ci}$	$Q_c$ <sup>1</sup>

<sup>1</sup> Variables 1 to 3 represent measurements without noise, this study utilizes variables 4 to 10.

The training and testing sets were collected from the CSTR simulation model during a 1200-second run, with a sampling rate of one sample per second. Each testing set initiates from a fault-free state and introduces faults after running for 200 seconds. Four fault scenarios were employed to evaluate the effectiveness of proposed method, including two input subspace faults and two output subspace faults, with detailed fault information provided in Table 4.

Table 4. Fault detailed information.

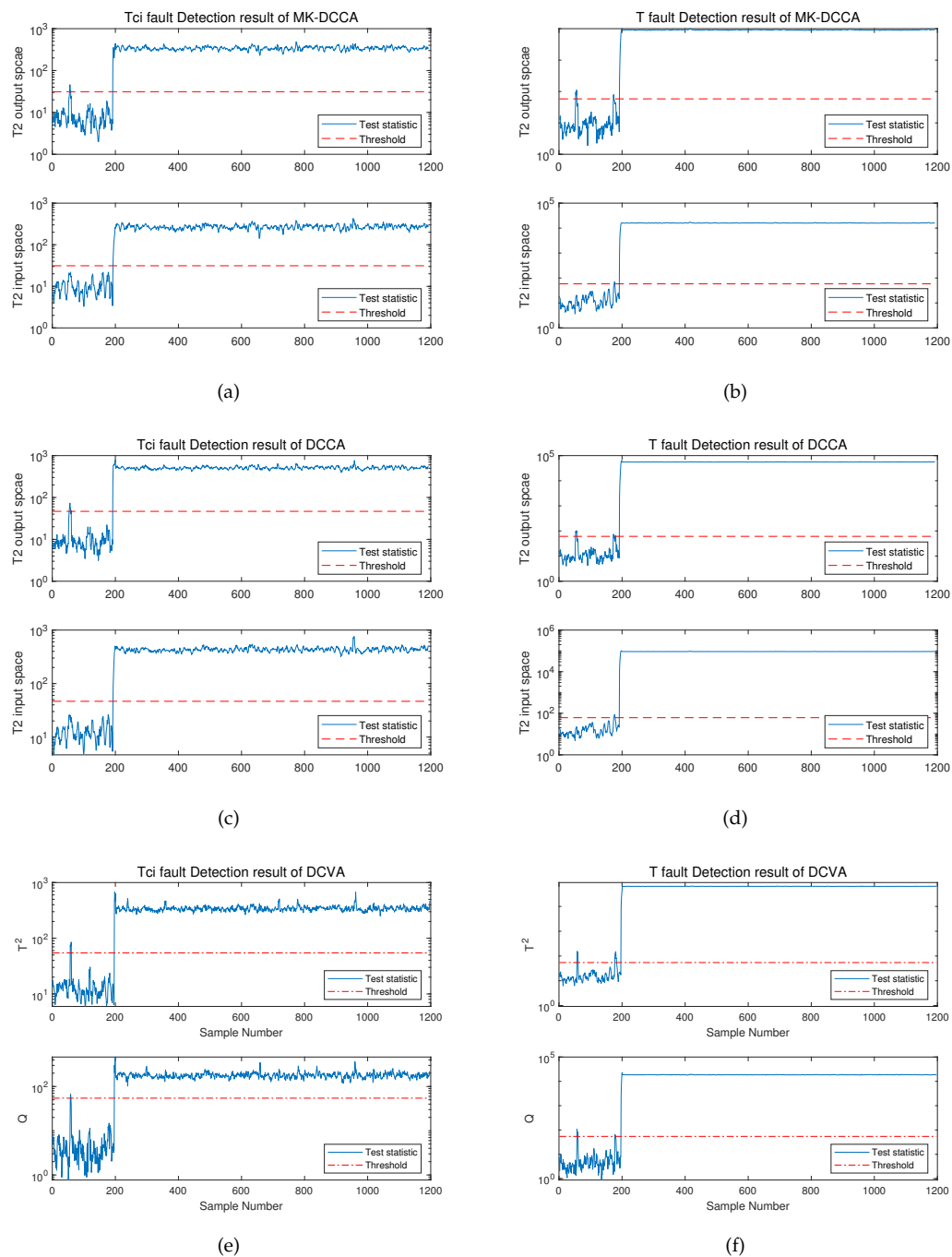
Fault index	Simulated fault scenario	Fault variables	Fault Category	Introduction Time(s)	Associated Subspace
Fault 1	Feed valve malfunction	$C_i$	Incipient	200	Input
Fault 2	High Coolant Temperature	$T_{ci}$	Abrupt	200	Input
Fault 3	Coolant leakage	$Q_c$	Incipient	200	Output
Fault 4	High Reactor Temperature	$T$	Abrupt	200	Output
Fault 5	Both 1 and 3	$C_i, Q_c$	Incipient	200	/

4.2.2. Process Monitoring

In this section, we initiate comparative experiments for the relatively easily detectable abrupt faults, Fault 2 and 4. Subsequently, experiments are conducted on Fault 1 and 3 (incipient faults), followed by the simultaneous introduction of both incipient faults of Fault 5 for detection. The methods employed include MK-DCCA, DCCA, and DCVA.

The process monitoring experimental results for Fault 2 and 4 are illustrated in Figure 4.





**Figure 4.** Monitoring results for Fault 2 and 4 by three methods:(a)Monitoring results of MK-DCCA approach for Fault 2.(b)Monitoring results of MK-DCCA approach for Fault 4.(c)Monitoring results of DCCA approach for Fault 2.(d)Monitoring results of DCCA approach for Fault 4.(e)Monitoring results of DCVA approach for Fault 2.(f)Monitoring results of DCVA approach for Fault 4.

From the graphs, it's evident that all three methods can provide early warnings for both Fault 2 and Fault 4 as they occur. However, in terms of false alarm rate, MK-DCCA exhibits the lowest, followed by DCCA, and DCVA performs the poorest. To facilitate better comparison of the three methods, Table 5 presents the detailed information of their monitoring performance indicators. The results demonstrate that both MK-DCCA and DCCA methods achieve a fault detection rate of 100% for both faults 2 and 4, surpassing DCVA method. Meanwhile, both MK-DCCA and DCCA methods exhibit no false alarms in the monitoring of fault 2, whereas DCVA method achieves the best performance with a false alarm rate

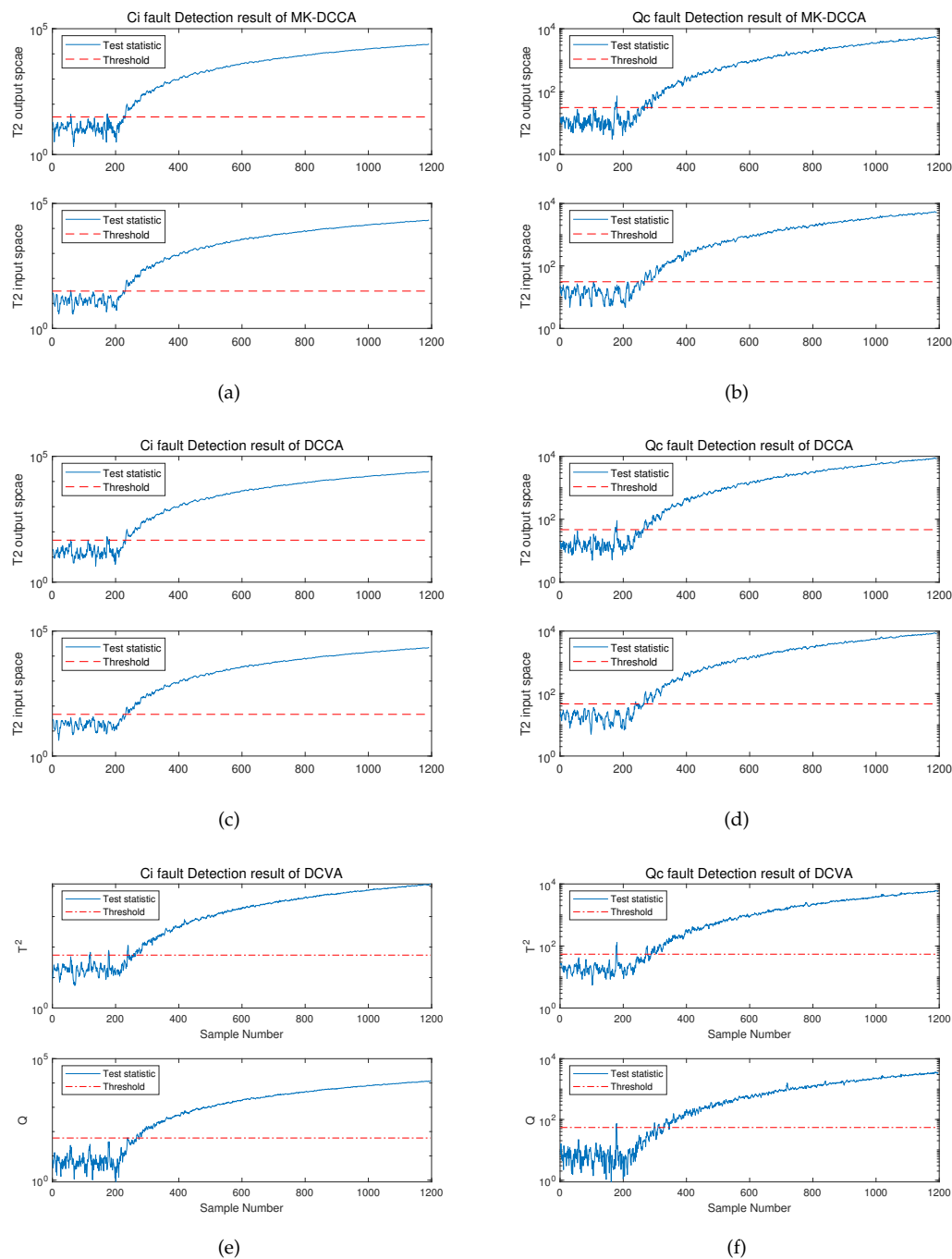
of 1.05% based on the Q criterion. In experiments for fault 4, MK-DCCA method similarly achieves the lowest false alarm rate of 1.05%, followed by DCCA method at 2.62%, and the DCVA method performs the least favorably with a rate of 3.66%. Furthermore, in terms of missed detection rate, both CCA-based methods maintained the lowest 0 missed detection rate in experiments for faults 2 and 4, while the DCVA method had a rate of 0.5%. Lastly, in comparison to the CVA-based method, the fault detection time of the CCA-based method is slightly reduced in both fault scenarios.

**Table 5.** Performance indicators for monitoring of Fault 2 and 4 using different methods.

Fault Index	Method	Statistical Metrics	FDR(%)	FAR(%)	MDR(%)	FDT(s)
Fault 2	MK-DCCA	$T_{in}^2$	99.8	0	0.2	202
		$T_{out}^2$	100	0	0	200
	DCCA	$T_{in}^2$	99.9	0	0.1	201
		$T_{out}^2$	100	3.14	0	200
	DCVA	Q	99.5	1.05	0.5	205
		$T^2$	99.5	2.1	0.5	205
Fault 4	MK-DCCA	$T_{in}^2$	100	1.05	0	200
		$T_{out}^2$	100	2.62	0	200
	DCCA	$T_{in}^2$	100	2.62	0	200
		$T_{out}^2$	100	6.81	0	200
	DCVA	Q	99.5	3.66	0.5	205
		$T^2$	99.5	5.24	0.5	205

To conclude, in terms of abrupt fault detection, the proposed MK-DCCA method performs the best, followed by the DCCA method, and finally the CVDA method. Therefore, applying the proposed method for detecting abrupt faults in industrial process systems is justified in this study.

Furthermore, the study further compared and analyzed the monitoring performance of the proposed method and the comparative methods in the incipient fault scenarios of Fault 1 and 3. The process monitoring experimental results for Fault 1 and 3 are depicted in Figure 5.



**Figure 5.** Monitoring results for Fault 1 and 3 by three methods: (a) Monitoring results of MK-DCCA approach for Fault 1. (b) Monitoring results of MK-DCCA approach for Fault 3. (c) Monitoring results of DCCA approach for Fault 1. (d) Monitoring results of DCCA approach for Fault 3. (e) Monitoring results of DCVA approach for Fault 1. (f) Monitoring results of DCVA approach for Fault 3.

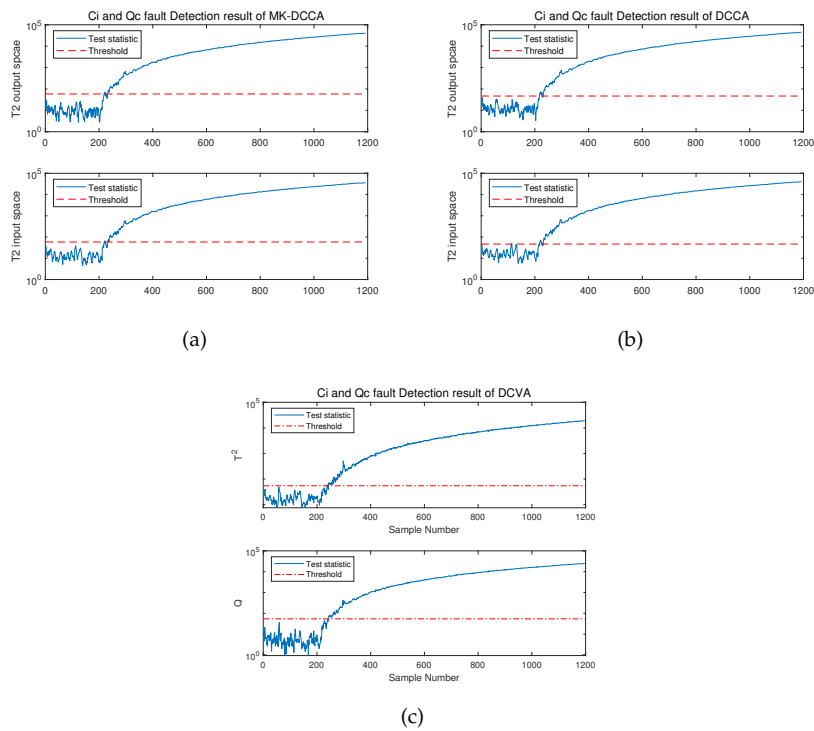
The figures clearly illustrate that the MK-DCCA, DCCA, and DCVA methods all exhibit the ability to provide alerts after a certain time period following the occurrence of faults, and simultaneously, they are capable of predicting the trend of fault progression. However, in comparison, the MK-DCCA method still performs better and more comprehensively, and detailed analysis of performance indicators can be found in Table 6. It is evident that the MK-DCCA method attains fault detection rates of 96.304% and 93.007% for the monitoring of fault 1 and fault 3, respectively, surpassing the FDR values of the DCCA and DCVA methods. Concurrently, in the fault scenarios of fault 1 and 3, the

CCA-based approach can attain lower false alarm rates than the CVA-based approach. Additionally, in terms of missed detection rates, the MK-DCCA method maintains the lowest rates in the monitoring of both faults, at 3.7% and 7.0%, respectively. As for fault detection time, the CCA-based method significantly reduces detection time compared to the CVA-based method, although the DCCA method also exhibits shorter detection time than the MK-DCCA method, it comes at the cost of slightly higher false alarm rates, making the performance of MK-DCCA more satisfactory in comparison.

**Table 6.** Performance indicators for monitoring of Fault 2 and 4 using different methods.

Fault Index	Method	Statistical Metrics	FDR(%)	FAR(%)	MDR(%)	FDT(s)
Fault 1	MK-DCCA	$T_{in}^2$	96.204	5.24	3.80	223
		$T_{out}^2$	96.304	1.57	3.70	225
	DCCA	$T_{in}^2$	96.004	0	4.0	231
		$T_{out}^2$	96.004	3.14	4.0	231
	DCVA	$Q$	92.607	0	7.40	237
		$T^2$	93.506	3.14	6.50	236
Fault 3	MK-DCCA	$T_{in}^2$	93.007	0	7.0	246
		$T_{out}^2$	92.408	3.14	7.60	259
	DCCA	$T_{in}^2$	92.907	0	7.10	238
		$T_{out}^2$	92.907	2.62	7.10	260
	DCVA	$Q$	87.013	1.57	12.99	297
		$T^2$	91.009	2.09	8.99	271

In summary, in experiments for detecting incipient faults, MK-DCCA method proposed in this paper continues to exhibit the best performance, followed by DCCA method, and finally CVDA method. Accordingly, the performance of proposed method has been demonstrated to be satisfactory in both abrupt fault and incipient fault detection scenarios. Additionally, in order to align the research more closely with complex real-world application scenarios, faults 1 and 3 were simultaneously introduced into the CSTR system, referred to as fault 5. The detailed results of the fault detection experiment for fault 5 are shown in Figure 6, and the specific values of the related performance indicators are presented in Table 7.



**Figure 6.** Monitoring results for Fault 5 by three methods:(a)Monitoring results of MK-DCCA.(b)Monitoring results of DCCA.(c)Monitoring results of DCVA.

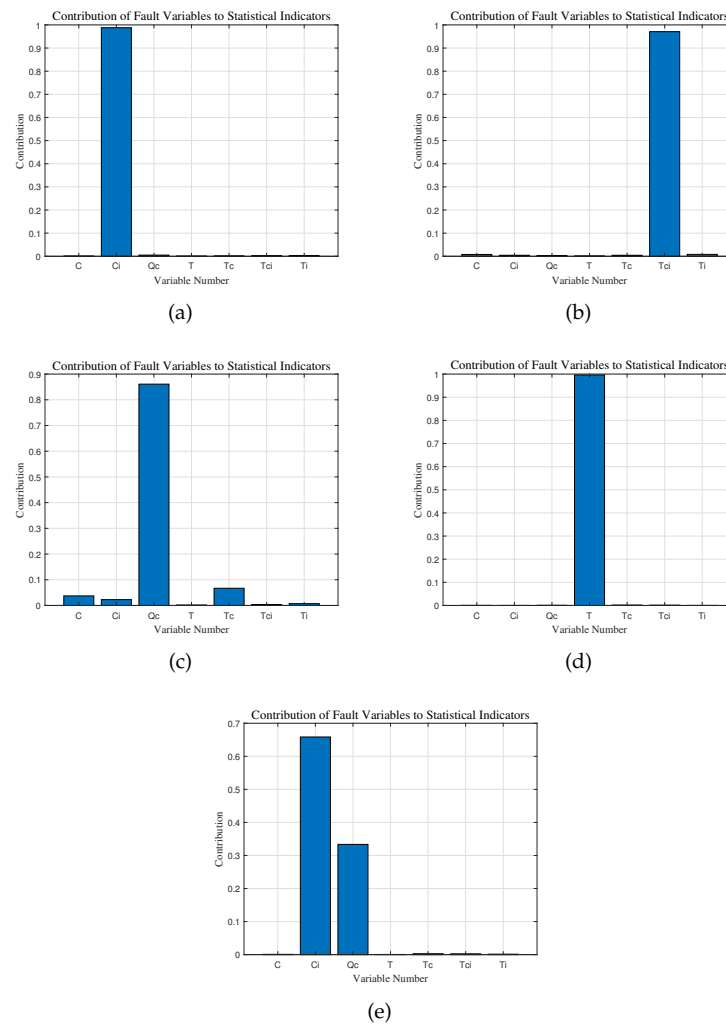
**Table 7.** Performance indicators for monitoring of Fault 5 using different methods.

Method	Statistical Metrics	FDR(%)	FAR(%)	MDR(%)	FDT(s)
MK-DCCA	$T_{in}^2$	97.203	1.04	2.80	215
	$T_{out}^2$	97.403	0	2.60	216
DCCA	$T_{in}^2$	97.103	1.05	2.90	215
	$T_{out}^2$	97.303	0	2.70	216
DCVA	$Q$	94.805	0	5.19	243
	$T^2$	95.105	0	4.90	237

It is evident that the CCA-based method holds a notable advantage over the CVA-based method, exhibiting a higher detection rate by around 2%, a lower miss detection rate by approximately %, and a reduction in detection time by about 20 seconds. Simultaneously, in the comparison between MK-DCCA and DCCA methods, the former prevails with a slight advantage in detection rate and miss detection rate. This reasserts the excellence of proposed MK-DCCA method and its feasibility in intricate application environments.

4.2.3. Fault Identification

In the previous section, the fault detection performance of the proposed method has been validated. In this section, the main emphasis lies in the analysis fault indentification capability of the method. Since it has been demonstrated in Section 4.1.3 that the  $T^2$ -based contribution identification accuracy is the highest, in experiments of this section, only the  $T^2$ -based contribution is used for fault identification. The contribution plots for Fault 1 to 5 are shown in Figure 7.



**Figure 7.** Monitoring results for Fault 5 by three methods:(a) Contribution plot of variables for fault1. (b) Contribution plot of variables for fault2. (c) Contribution plot of variables for fault3. (d) Contribution plot of variables for fault4. (e) Contribution plot of variables for fault5.

It can be observed that in Fault 1, the variable  $C_i$  contributes 98.66% to the statistical indicator, with the fact that  $C_i$  is the actual fault variable; in Faults 2, 3 and 4, the contributions of the fault variables  $T_{ci}$ ,  $Q_c$ , and  $T$  are 97.52%, 86.32%, and 99.32%, respectively, far exceeding other variables; in Fault 5, which involves the fault variables  $C_i$  and  $Q_c$ , their contributions are 64.3% and 34.84%, respectively, also significantly higher than other variables. The above findings demonstrate that the fault identification approach adopted in this study exhibits a satisfactory identification accuracy, effectively identifying and locating faults with precision.

## 5. Conclusions

This paper emphasizes the importance of detecting incipient faults in process industrial systems and extends the widely recognized process monitoring method, CCA, to make it more suitable for early detection of incipient faults. By incorporating time parameters, the method gains the ability to handle system dynamics. The inclusion of kernel methods endows the method with the capability to handle nonlinear data. In the selection of kernel functions, a weighted combination of RBF and polynomial kernels is chosen, allowing the kernel function to possess both good interpolation and extrapolation capabilities. Based on the aforementioned work, this paper proposes an MK-DCCA fault diagnosis method, and its generalization performance is validated on a randomly generated dataset,

then comparative experiments are conducted on the CSTR Simulink model. The results demonstrate the superiority of the proposed method in fault detection, especially in the case of incipient faults over the DCCA and DCVA methods.

Nevertheless, this study has certain limitations. The method requires a considerable number of parameters, and its performance is somewhat dependent on the selection of these parameters. The calculation of thresholds is relatively inflexible, leading to limited adaptability to different monitoring objects. Future research will aim to address these issues by exploring adaptive parameter selection and threshold computation.

**Author Contributions:** Conceptualization, J.W. and M.Z.; methodology, J.W. and M.Z.; validation, J.W. and L.C.; formal analysis, J.W.; investigation, J.W. and L.C.; resources, M.Z.; data curation, L.C.; writing—original draft preparation, J.W.; writing—review and editing, M.Z. and J.W.; visualization, J.W.; supervision, M.Z.; project administration, M.Z.; funding acquisition, M.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China, (Grant No. 62003106), Provincial Natural Science Foundation of Guizhou Province, China (Grant No. ZK (2021) 321. and [2017]5788).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data that support the findings of this study are available from the corresponding author upon reasonable request

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Tang, M.; Yang, C.; Gui, W. Fault detection based on cost-sensitive support vector machine for alumina evaporation process. *Control Engineering of China* **2011**, *18*, 645–649.
2. Qin, S.J. Survey on data-driven industrial process monitoring and diagnosis. *Annual reviews in control* **2012**, *36*, 220–234.
3. Wise, B.M.; Gallagher, N.B. The process chemometrics approach to process monitoring and fault detection. *Journal of Process Control* **1996**, *6*, 329–348.
4. Tessier, J.; Duchesne, C.; Tarcy, G.; Gauthier, C.; Dufour, G. Analysis of a potroom performance drift, from a multivariate point of view. LIGHT METALS-WARRENDALE-PROCEEDINGS-. TMS, 2008, Vol. 2008, p. 319.
5. Abd Majid, N.A.; Taylor, M.P.; Chen, J.J.; Stam, M.A.; Mulder, A.; Young, B.R. Aluminium process fault detection by multiway principal component analysis. *Control Engineering Practice* **2011**, *19*, 367–379.
6. Ding, S.X.; Yin, S.; Peng, K.; Hao, H.; Shen, B. A novel scheme for key performance indicator prediction and diagnosis with application to an industrial hot strip mill. *IEEE Transactions on Industrial Informatics* **2012**, *9*, 2239–2247.
7. Harkat, M.F.; Mansouri, M.; Nounou, M.N.; Nounou, H.N. Fault detection of uncertain chemical processes using interval partial least squares-based generalized likelihood ratio test. *Information Sciences* **2019**, *490*, 265–284.
8. Ruiz-Cárcel, C.; Cao, Y.; Mba, D.; Lao, L.; Samuel, R. Statistical process monitoring of a multiphase flow facility. *Control Engineering Practice* **2015**, *42*, 74–88.
9. Pilario, K.E.S.; Cao, Y. Canonical variate dissimilarity analysis for process incipient fault detection. *IEEE Transactions on Industrial Informatics* **2018**, *14*, 5308–5315.
10. Pilario, K.E.S.; Cao, Y.; Shafiee, M. Mixed kernel canonical variate dissimilarity analysis for incipient fault monitoring in nonlinear dynamic processes. *Computers & Chemical Engineering* **2019**, *123*, 143–154.
11. Pilario, K.E.S.; Cao, Y.; Shafiee, M. Incipient Fault Detection, Diagnosis, and Prognosis using Canonical Variate Dissimilarity Analysis. In *Computer Aided Chemical Engineering*; Elsevier, 2019; Vol. 46, pp. 1195–1200.
12. Chen, Z.; Ding, S.X.; Zhang, K.; Li, Z.; Hu, Z. Canonical correlation analysis-based fault detection methods with application to alumina evaporation process. *Control Engineering Practice* **2016**, *46*, 51–58.
13. Chen, Z.; Deng, Q.; Zhao, Z.; Tang, P.; Luo, W.; Liu, Q. Application of just-in-time-learning CCA to the health monitoring of a real cold source system. *IFAC-PapersOnLine* **2022**, *55*, 23–30.



14. Chen, Z.; Zhang, K.; Ding, S.X.; Shardt, Y.A.; Hu, Z. Improved canonical correlation analysis-based fault detection methods for industrial processes. *Journal of Process Control* **2016**, *41*, 26–34.
15. Gao, L.; Li, D.; Yao, L.; Gao, Y. Sensor drift fault diagnosis for chiller system using deep recurrent canonical correlation analysis and k-nearest neighbor classifier. *ISA transactions* **2022**, *122*, 232–246.
16. Li, X.; Wang, Y.; Tang, B.; Qin, Y.; Zhang, G. Canonical correlation analysis of dimension reduced degradation feature space for machinery condition monitoring. *Mechanical Systems and Signal Processing* **2023**, *182*, 109603.
17. Chen, Z.; Liang, K. Canonical correlation analysis-based fault diagnosis method for dynamic processes. In *Fault Diagnosis and Prognosis Techniques for Complex Engineering Systems*; Elsevier, 2021; pp. 51–88.
18. Schölkopf, B.; Smola, A.; Müller, K.R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation* **1998**, *10*, 1299–1319.
19. Smola, A.; Ovári, Z.; Williamson, R.C. Regularization with dot-product kernels. *Advances in neural information processing systems* **2000**, *13*.
20. Cristianini, N.; Shawe-Taylor, J. *An introduction to support vector machines and other kernel-based learning methods*; Cambridge university press, 2000.
21. Negiz, A.; Çlinar, A. Statistical monitoring of multivariable dynamic processes with state-space models. *AIChE Journal* **1997**, *43*, 2002–2020.
22. Li, X.; Mba, D.; Diallo, D.; Delpha, C. Canonical variate residuals-based fault diagnosis for slowly evolving faults. *Energies* **2019**, *12*, 726.
23. Jiang, B.; Huang, D.; Zhu, X.; Yang, F.; Braatz, R.D. Canonical variate analysis-based contributions for fault identification. *Journal of Process Control* **2015**, *26*, 17–25.
24. Jordaan, E.M. Development of robust inferential sensors: Industrial application of support vector machines for regression. **2004**.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.